Topic Cube: OLAP of Text Data

Duo Zhang, Jiawei Han, Chengxiang Zhai
Department of Computer Science, University of Illinois at Urbana Champaign
{*dzhang22, hanj, czhai*}@cs.uiuc.edu

Abstract

In this poster, we present a new concept called *Topic Cube*, which is the first attempt of extending OLAP technology for exploring text data. The power of a topic cube is to enable a user to analyze a large set of documents in different contexts and topics with multiple granularity.

In data warehouse system, OLAP (Online Analytical Analysis) has been well used for analyzing multidimensional data. Traditional data cubes are capable of efficiently computing Average or Sum of numeric data. However, no previous OLAP technique has been designed for exploring nonnumeric data like documents. For example, with a traditional data cube a supermarket manager can easily get the number of sales of one product during a certain period, but what she won't get is why the number increases or decreases compared with other periods. Nowadays, topic modeling is one of the most popular methods for mining and analyzing text data, which summarizes a set of documents in topics. The combination of topic modeling methods with a data cube will definitely enrich the functions of traditional data cubes. For instance, in the previous example, if we use topic modeling methods to summarize users' online reviews about products during different periods and store them in a data cube, a supermarket manager will easily see what users are talking about the products and thus get the idea of why the sales increase or decrease.

In our study, we formally define the concept of topic cube. Generally speaking, a topic cube is constructed based on a text database and a hierarchical topic tree. The main difference between a traditional cube and a topic cube is the topic dimension added in a topic cube. Drill-down and roll-up operations are defined along the topic dimension, which allow users to analyze the topics from different granularities. One big issue in our study is how to efficiently materialize a topic cube. Instead of exhaustively constructing a topic cube from every scratch, we propose a heuristic method that efficiently constructs a topic cube. Another issue is how to reduce storage cost of topic cubes. We proposed an approximate method which only stores top *k* words for each topic in each cell. The ASRS (Aviation Safety Reporting System) database is used as our test data set to evaluate the proposed methods. In ASRS database, we regard the 'Anomaly Event' field as a topic dimension. The topic cube we constructed based on ASRS aims at assisting experts to analyze safety issues during certain contexts, such as time, location, and environment. Experimental results show that our materialization methods are much more efficient than a baseline method.